Title: Watermark Estimation Through Detector Observations
Author: Ton Kalker
Conference: Benelux Signal Processing Symposium 98
Date and location: March 98, Leuven, Belgium

Form SF298 Citation Data

Report Date *("DD MON YYYY")* 01031998	**Report Type** N/A	**Dates Covered (from... to)** *("DD MON YYYY")*

Title and Subtitle Watermark Estimation Through Detector Observations	**Contract or Grant Number**
	Program Element Number
Authors	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) IATAC Information Assurance Technology Analysis Center 3190 Fairview Park Drive Falls Church VA 22042	**Performing Organization Number(s)**
Sponsoring/Monitoring Agency Name(s) and Address(es)	**Monitoring Agency Acronym**
	Monitoring Agency Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes	
Abstract	
Subject Terms	
Document Classification unclassified	**Classification of SF298** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** unlimited
Number of Pages 6	

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** 3/1/98	**3. REPORT TYPE AND DATES COVERED** Report

4. TITLE AND SUBTITLE Watermark Estimation Through Detector Observations	**5. FUNDING NUMBERS**
6. AUTHOR(S) Not provided	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Information Assurance Technology Analysis Center (IATAC) 3190 Fairview Park Drive Falls Church, VA 22042	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Defense Technical Information Center DTIC-AI 8725 John J. Kingman Road, Suite 944	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT	**12b. DISTRIBUTION CODE** A

13. ABSTRACT *(Maximum 200 Words)*

A watermark is a perceptually unobtrusive signal embedded in an image, an audio or video clip, or any other other multimedia asset. Its purpose is to be a label which is holographically attached to the content. Moreover, it can only be removed by malicious and deliberate attacks (without a great loss of content quality) if some secret parameter K is known. In contrast, a watermark should be readily detectable by electronic means. This implies that electronic watermark detection is only feasible if the watermark detector is aware of the secret K. In many watermarking business scenarios the watermark detector will be available to the public as a black box D. The following question is therefore justi_ed: can the secret K be deduced from the operation of the black box D? And if yes, what is the complexity of this process? In this paper
we will address this issue for the watermarking method PatchWork [1].

14. SUBJECT TERMS Watermark, Information Security	**15. NUMBER OF PAGES**
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

Watermark Estimation Through Detector Observations

Ton Kalker

Philips Research Eindhoven
Prof. Holstlaan 4, 5656 AA Eindhoven
The Netherlands

31-40-2743839 (tel)
31-40-2744675 (fax)
kalker@natlab.research.philips.com (email)

Abstract

A watermark is a perceptually unobtrusive signal embedded in an image, an audio or video clip, or any other other multimedia asset. Its purpose is to be a label which is holographically attached to the content. Moreover, it can only be removed by malicious and deliberate attacks (without a great loss of content quality) if some secret parameter K is known. In contrast, a watermark should be readily detectable by electronic means. This implies that electronic watermark detection is only feasible if the watermark detector is aware of the secret K. In many watermarking business scenarios the watermark detector will be available to the public as a black box $\mathbf{D}$. The following question is therefore justified: can the secret K be deduced from the operation of the black box $\mathbf{D}$? And if yes, what is the complexity of this process? In this paper we will address this issue for the watermarking method PatchWork [1].

1 Introduction

Watermarking is a fundamental enabling technology for the distribution of digital multimedia (MM) content. At present it is very easy to distribute and copy digital multimedia content. Without any special precautions the content generation and distribution industry will be very reluctant to publish in the digital domain. The slow introduction of the new Digital Versatile Disk (DVD) format bears witness to this tendency.

Digital watermarking is a technical solution to the copyright problem. In its basic form a digital watermark W is a *small* signal added to MM content. The watermark W carries sufficient data to ensure proper copyright verification. Due to its intended purpose a watermark should be unobtrusive (i.e. no perceptible degradation of the quality is allowed), easily detectable by dedicated software or hardware and very difficult to remove by malicious and deliberate attacks.

It is essential to distinguish two types of applications of watermarking technology. In the first type of application all content can be enforced to contain a watermark. A typical example is given by (images on) bank notes and smart cards. It is not sufficient for a pirate to remove the watermark (i.e. reconstruct the original content), but he will actually have to insert a watermark which contains false copyright information. By relying on cryptographic methods the complexity of this type of attack can be made arbitrarily large.

In the second type of application watermarking cannot be enforced. A typical example is given by film content on DVD. The film industry can enforce watermarks on commercial digital video, but it cannot enforce watermarking of home videos. Therefore DVD players will have to accept both watermarked (i.e. copyright protected) and unwatermarked content. This implies that it is sufficient for a pirate to remove a watermark from a commercial video (i.e. make a good estimate of the unwatermarked original, also referred to as *unzigning* [2]) in order to invalidate the copyright protection mechanism of DVD. In this paper we will focus on this type of non-watermark-enforced application.

In particular we will study the security risk associated to the availability of a watermark detector. This applies for example to DVD, where a copyright system based on watermarking will imply a watermark detector in every single DVD player. We will assume that a pirate has a general knowledge of the watermark embedding and detection scheme, but not of the associated secrets such as keys K or noise patterns W. The key question addressed in this paper is whether or not the availability of a detector $\mathbf{D}$ allows the retrieval of a sufficient amount of secret information to unzign watermarked material.

It is obvious very difficult to study this situation in full generality. Therefore we will confine ourselves to PatchWork, a specific still image watermarking scheme introduced by Bender *et al.* [1]. Although PatchWork is a relatively simple scheme, it is prototypical for a large class of watermarking schemes.

This paper is organized as follows. In Section 2 we recall the PatchWork watermarking scheme. In Section 3 we present a method for retrieving the secret pattern W associated to PatchWork. In Section 4 this method is experimentally validated. Section 5 summarizes the paper.

2 The PatchWork Watermarking Scheme

We recall the watermarking procedure PatchWork as introduced in [1]. Given an original image $X = \{x_i\}$ of size $N_1 \times N_2$ the watermark casting procedure of PatchWork starts with choosing a secret K. Depending on this key K the set of image points is partitioned into three sets A, B and C, where A and B are of equal size. The watermarked image $Y = \{y_i\}$ is now obtained by increasing the luminance values of the pixels in A with k, decreasing the luminance values of B with k and leaving the pixels in C unchanged. The value of k depends on the desired robustness of the watermark (better with larger k) and the desired imperceptibility (better with smaller k).

The detection process starts with modifying a suspect image $Z = \{z_i\}$ to have zero mean. Without loss of generality we may therefore assume that $\mathrm{mean}(Z) = 0$. Subsequently the sets A, B and C are derived from the secret key K (to which the detector has access!). The detector then computes the decision value d as

$$d = \frac{1}{N}\left(\sum_{i\in A} - \sum_{i\in B}\right), \tag{1}$$

where $N = N_1 N_2$ is the total number of pixels. It is not difficult to see that for an unwatermarked image the expected value of d is equal to 0. For a watermarked image however the expected value of d is equal to $2kM/N$, where M is the size of A. Invoking some basic statistical theory it not difficult to derive that the standard deviation σ_d of the decision variable d is given by

$$\sigma_d \approx \frac{s_Z\sqrt{2kM}}{N}, \tag{2}$$

where the standard deviation s_Z of the image Z is defined by

$$s_Z = \sqrt{\frac{\sum_i z_i^2}{N}} \tag{3}$$

Given a threshold parameter T, PatchWork decides that a suspect image Z is watermarked if and only if the computed decision variable d is above $T\sigma_d$. In particular it follows that we have to constrain $\sqrt{2kM}$ to be larger than Ts_Z in the casting procedure. Otherwise a watermarked image will not be recognized as such by the detector.

PatchWork can be reformulated into a somewhat more abstract mathematical setting. The watermark embedding procedure can be described as adding a noise sequence W to an image X. The noise sequence $W = \{w_i\}$ is determined by the secret key K. Moreover, W is $\{-1, 0, 1\}$-valued and *DC-free*, i.e $\sum_i w_i = 0$. The watermarked image Y is then obtained as $Y = X + kW$. The detection procedure can be described as the computation of a thresholded correlation value. Firstly, a suspect image Z is correlated with the secret noise pattern W to obtain a decision value d,

$$d = \langle Z, W\rangle = \frac{1}{N}\sum_i z_i w_i. \tag{4}$$

This value d is normalized to unit standard deviation by dividing by σ_d (see Equation 2) and then compared with the threshold T. If the outcome is larger than T the image is said to be watermarked and otherwise it is not. Summarizing the above, the detection process can be described by the following pseudo-code.

```
$Z := Z - \mathrm{mean}(Z)$;
$Z := Z/\,\mathrm{std}(Z)$;
$d := (\sum_{i=0}^{N-1} z_i w_i)/\sqrt{2M}$;
**if** $d < T$
    **return** $-1$;
**else**
    **return** $1$;
**fi**
```

It should be obvious from the preceding that knowledge of W allows the removal of watermarks (*unzigning*). For applications in which $\mathbf{D}$ is publicly available it is therefore essential that $\mathbf{D}$ does not reveal any information about W. This is exactly the security issue addressed in this paper.

Assuming that an attacker has access to a detector $\mathbf{D}$ and a watermarked image X_0 we analyze the security of the PatchWork scheme. In particular, referring to Section 1, we pose the questions of how much knowledge can be obtained about W and what effort is needed to obtain this information. The following sections will show that it is relatively easy to make an estimation of W.

3 The PatchWork Attack

The simplest method to obtain information about W is by brute force trial and error. Starting with a fixed (unmarked) image X we add a DC-free sequence V to X, and we check if the resulting image is watermarked or not (using the detector $\mathbf{D}$). If the outcome is negative we choose another sequence V and continue this process until the detector indicates that we have found a watermarked image. The sequence V is then our estimation of W. It is not difficult to see that this process is exponential in the number of pixels N, and is therefore infeasible in practice.

In this paper we propose an attack which, for any desired accuracy, is polynomial in N. The basic idea of the attack is to fix an image X_1 and offer small perturbations of X_1 to the detector $\mathbf{D}$. By observing the output of $\mathbf{D}$, information about W is obtained. It is obvious that nothing can be learned if the start image X_1 is far below the watermarking threshold T. Any small perturbation of X_1 will cause $\mathbf{D}$ to give out the decision "not watermarked". For the same reason nothing can be learned from a start image X_1 which is far above threshold. The first phase of the proposed attack therefore generates an image X_1 which is approximately at threshold. As we have indicated in the opening paragraph of this section, it is difficult to generate such an image from scratch. A watermarked image X_0 is therefore required as input. The signal X_1 can be obtained from X_0 by gradually reducing the image quality of X_0. Several methods can be applied for this purpose. One method iteratively replaces sample values of X_0 by the mean value of X_0. Another method consists of blurring X_0 using a decreasing pass band width. In the experiments presented in Section 4 we have chosen the former method as it has lower complexity. A geometrical interpretation of this first phase is represented in Figure 1. The curved dashed line in this figure represent the path traversed by X_0 when image quality is gradually decreasing.

In the second phase of the attack we perturb X_1 by adding DC-free $\{-k, +k\}$-valued noise sequences V and feeding $X_1 + V$ to $\mathbf{D}$. If the correlation between V and W is positive there is a slightly increased chance that $\mathbf{D}$ gives "watermarked" as output. If W and V are negatively correlated there is a slightly increased chance of $\mathbf{D}$ giving out "not watermarked". In the former case we take V as an approximation of W, in the latter case we take $-V$ as an approximation of W. These perturbations of X_1 are geometrically represented in Figure 1 as the shaded sphere around X_1. The lighter and darker shaded areas of this sphere indicate the negatively and positively correlated perturbations, respectively. By averaging over all intermediate estimations we obtain an estimate of W up to a positive scalar κ. The final approximation of W is obtained by scaling and quantizing to the domain $\{-1, 0, +1\}$.

When perturbing X_1 there is a choice in the strength k of the perturbations. If k is either too small or too large the variation in standard deviation of the perturbed images has a dominant effect on the outcome of the experiments. These cases can however easily be recognized by comparing the standard deviation of the intermediate estimation with the theoretical standard deviation for experiments on non-watermarked images (see step 8b in the pseudo-code below). If k is "just right" (quoting Goldilocks) then the measured standard deviation will be (significantly) larger than this theoretical standard deviation. In the pseudo-code below we obtain a proper value of k by monotonically increasing k, starting at the value $k = 1$.

A formal description of the attack is given below.

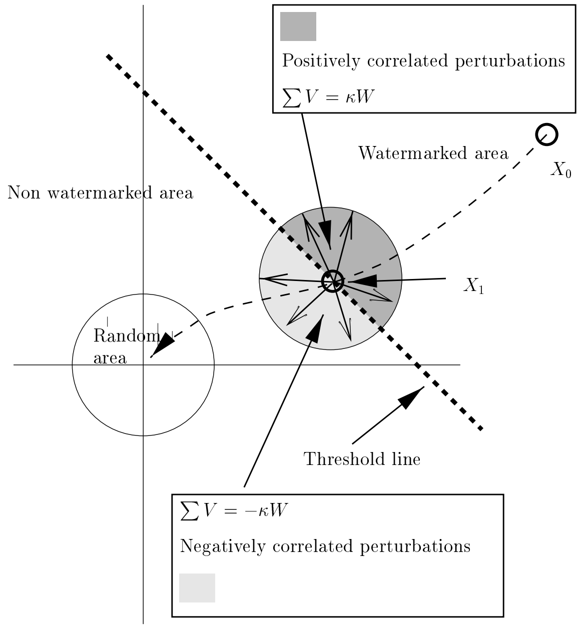

Figure 1: A geometrical interpretation of the watermark attack.

PatchWork Attack Pseudo-Code

```
Input:
  detector $\mathbf{D}$ ;
  watermarked image $X_0$ ;
  total number of iterations $K$ ;
Output:
  estimation $\hat{W}$ of watermark secret $W$ ;
1. Degrade $X_0$ to $X_1$ at threshold of detection.
2. (a) $k := 0$.
       { Initial perturbation parameter }
   (b) $kf := 0$.
       { $k$ too small }
3. $k := k + 1$.
   { Increment perturbation parameter }
4. (a) $\hat{W} = 0$.
       { Initialize watermark estimation }
   (b) $M = 0$.
       { Initialize loop counter }
5. (a) $M := M + 1$.
       { Increment loop counter }
   (b) $V = \mathrm{dcfree}()$.
       { Random DC-free $V$, $v_i \in \{-1, +1\}$ }
6. $d = \mathbf{D}(X_1 + k * V)$.
   { Perturb and record decision }
7. $\hat{W} := \hat{W} + d * V$.
   { Update $\hat{W}$ }
8. If $kf == 0$
   { Proper $k$ not yet found }
   (a) If $M < N$ go back to 5.
       { Too few iterations to make a decision }
   (b) Else
       i. If $\mathrm{std}(\hat{W}) \approx \sqrt{N}$ go back to 3
          { $k$ too small }
       ii. Else $kf := 1$.
          { $k$ large enough }
9. If $kf == 1$ and $M < K$ go back to 5
   { Not enough iterations }
10. $\hat{W} := \mathrm{quantize}(\hat{W})$.
   { Scale and round $\hat{W}$ to domain $\{-1, 0, +1\}$ }
```

It can be proved that this procedure recovers the secret W for any desired accuracy in $\mathcal{O}(N^2)$ experiments (i.e loop traversals). The full argumentation for the correctness of the presented attack and its complexity will be provided in a forthcoming paper. In [3] a slightly generalized attack will be presented which can also deal with a larger class of watermarking methods.

The basic idea of using perturbations for watermark hacking has first been presented in [4] and [5]. The attack methods in these papers are however mainly of a theoretical nature. The overall conclusion of both approaches however is the same: watermark detection schemes which are based on thresholded correlation can be cracked in $\mathcal{O}(N^2)$ experiments (i.e loop traversals).

4 Experiments

The attack described in Section 3 has experimentally been verified. An attack was simulated in MATLAB for a watermarked image[1] of size 128×128 (marked at 9 times standard deviation) and a software model of a watermark detector with $T = 7$. The chosen image size 128×128 is realistic in the sense that any practical watermarking scheme will use tiling with moderately sized tiles, usually smaller than 128×128. A watermark detector in such a system essentially computes correlation values on the basis of this tile size.

The watermarks sequence W used in this experiment has an average energy of 0.5, e.g. $M = N/4$ (see Section 2).

First an image at the threshold of detection was obtained by iteratively replacing sample values by the mean value of the image (see Figure 2). This threshold image was used for a number of experiments in which the quality of watermark retrieval was measured as a function of the perturbation strength k and the total number of iterations K. The results are given Figure 3. The quality

[1] The central part of the well known Lena image was chosen for this experiment.

of retrieval is measured in the percentage of watermarks bits which are correctly estimated. The number of iterations is measured in multiples of the number of samples $N = 128^2$. The figure clearly shows that the percentage of retrieval is easily above 90% for a moderated number of iterations. The figure also shows that the attack is reasonably insensitive to the exact value of k as long as k is not too small or too large. In Figure 4 the distribution of sample values of $\hat{W}$ before quantization is plotted for $K = 30 * N$ and $k = 5$. This figure clearly shows a distribution which is divided in 3 different regions. The left, middle and region correspond to the values which are quantized to -1, 0 and $+1$ respectively. From this figure one can also see that about half of the values in $\hat{W}$ is equal to 0.

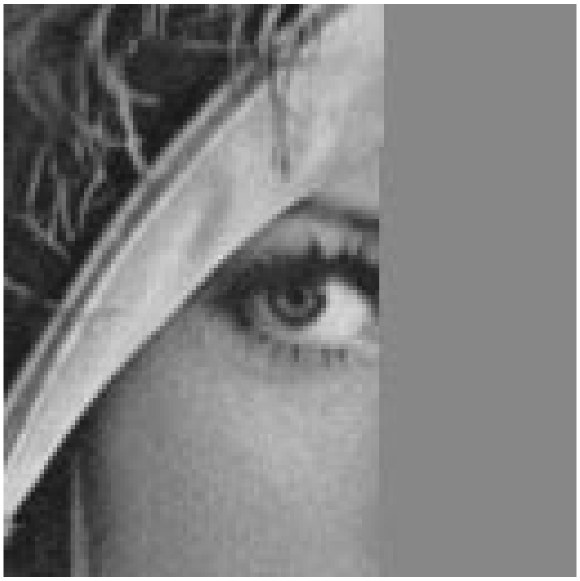

Figure 2: The (partial) Lena image at detection threshold.

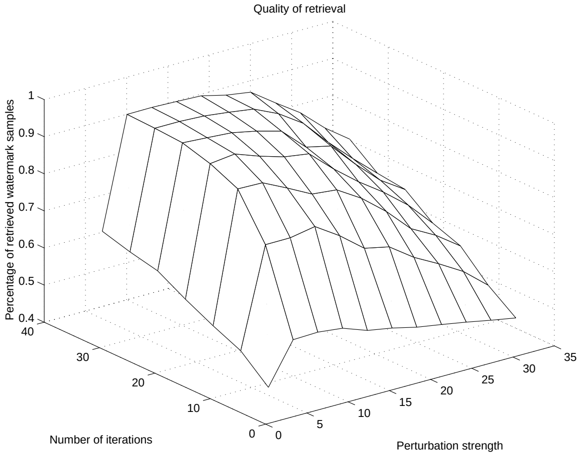

Figure 3: Retrieval percentage as a function of the perturbation strength k and the total number of iteration K.

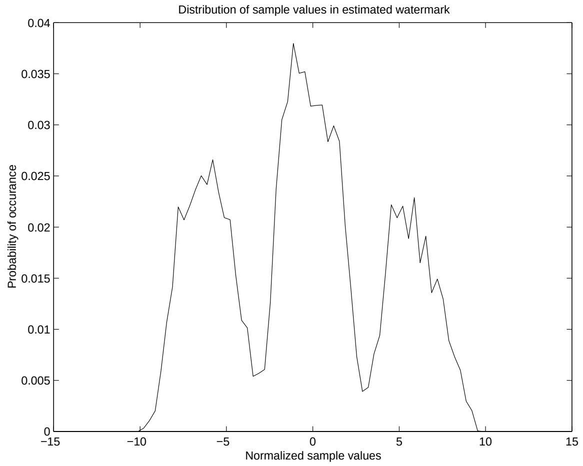

Figure 4: Distribution of sample values in $\hat{W}$ for $K = 30 * N$ and $k = 5$.

5 Conclusions

In this paper we have addressed the issue of watermark security based on the availability of a watermark detector and a single watermarked image. We extended the work started in [4] and [5] by presenting a simple attack method. We have shown that for the PatchWork the complexity of watermark retrieval is quadratic in the number of sample values. A slight extension of the theory shows that the same method is applicable to all correlation based watermark methods. This implies that watermarking methods based upon thresholded correlation are not suited for applications where watermarking cannot be enforced and the detector is publicly available (DVD).

References

[1] W. Bender, D. Gruhl, and N. Morimoto. Techniques for data hiding. In *Proceedings of the SPIE*, volume 2420, page 40, San Jose CA, USA, February 1995.

[2] The unzign web page. http://altern.org/watermark, 1997.

[3] T. Kalker, J.P Linnartz, and M. van Dijk. Watermark estimation through detector analysis. In *Proceedings of the ICIP*, Chicago, October 1998. Submitted.

[4] I.J. Cox and J.P.M.G. Linnartz. Public watermarks and resistance to tampering. In *Proceedings of the ICIP*, Santa Barbara, California, October 1997. Paper appears only in CD version of proceedings.

[5] J.P. Linnartz and M. van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In *Proceedings of the Workshop on Information Hiding*, Portland, April 1998. Submitted.